Multiple Regression Analysis for Life Expectancy at Birth: Python Application

Natalia Irena Gust-Bardon

January 2019

1. Problem Description

The purpose of this project is to obtain a prediction equation for the life expectancy at birth. The data set recorded mostly in 2015 consists of 102 observations (102 countries), one dependent variable (Life Expectancy) and the following 15 predictor variables representing four categories:

1. Environment
    - Forest area (% of land area)
    - Improved water source (% of population with access)
    - Renewable energy consumption (% of total final energy consumption)
2. Economic Development
    - Urban population (% of total)
    - GDP per capita
    - Services value added (% of GDP)
    - Exports of goods and services (% of GDP)
    - Developed economies (categorical variable)
    - Labor force participation rate, female (% of female population ages 15+)
    - Gini Index
3. Health
    - Health expenditure per capita (current US$)
    - Improved sanitation facilities (% of population with access)
    - Obesity (in %)
    - Prevalence of undernourishment (% of population)
4. Social Protection
    - Share of unemployed receiving regular periodic social security unemployment
    - Public social protection expenditure [excluding health care] as a percentage of GDP

This report presents the set of activities allowing me to build a multivariate regression model, including:

(a) checking for the violations of model assumptions;
(b) data transformation;
(c) variable selection techniques;
(d) incorporation of categorical explanatory variables;
(e) application of *F*-tests.

2. Investigation of the Data

The data set consists of fourteen quantitative variables and one categorical variable (Developed Economy).

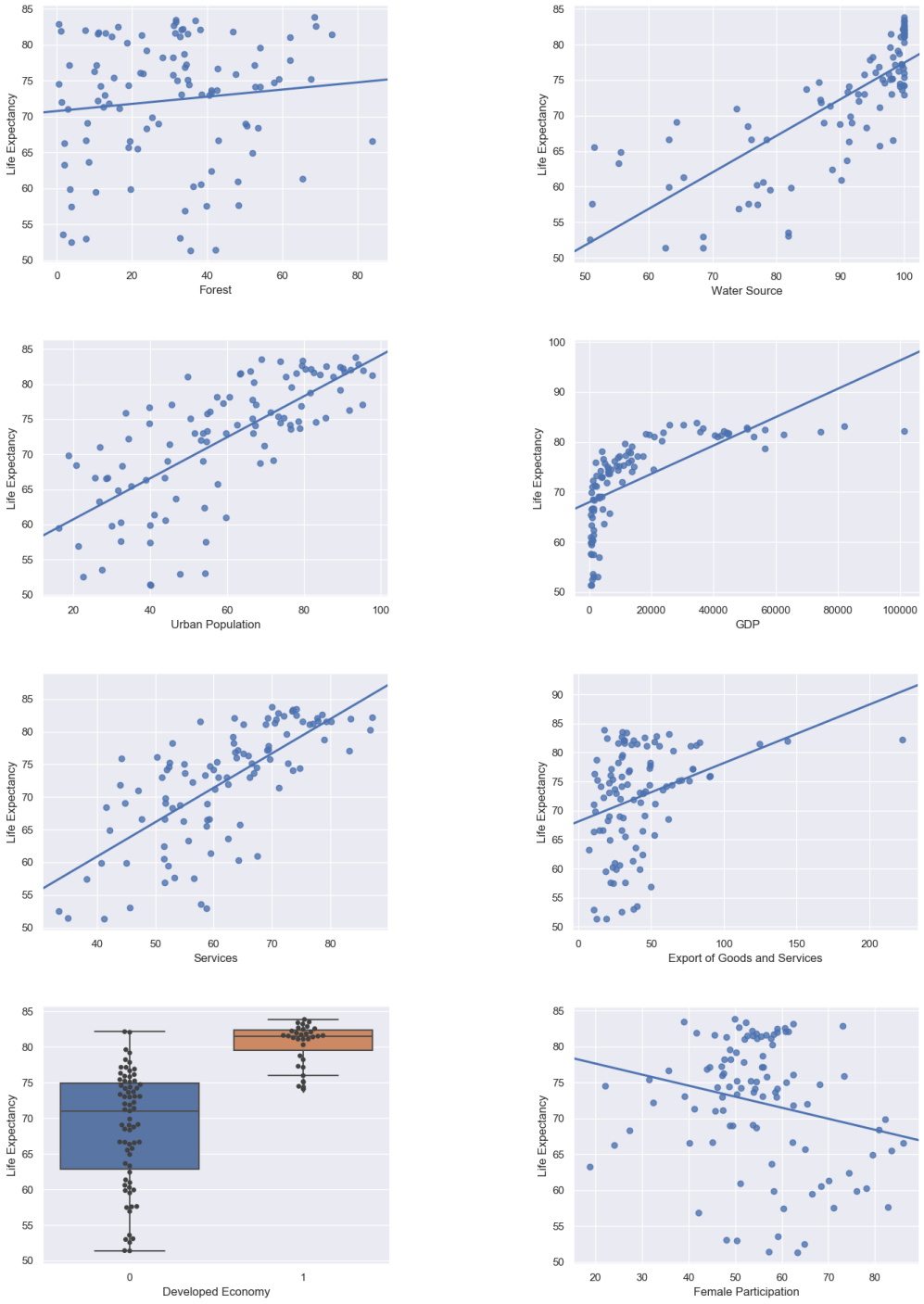

Figure 2.1 *Predictor variables*

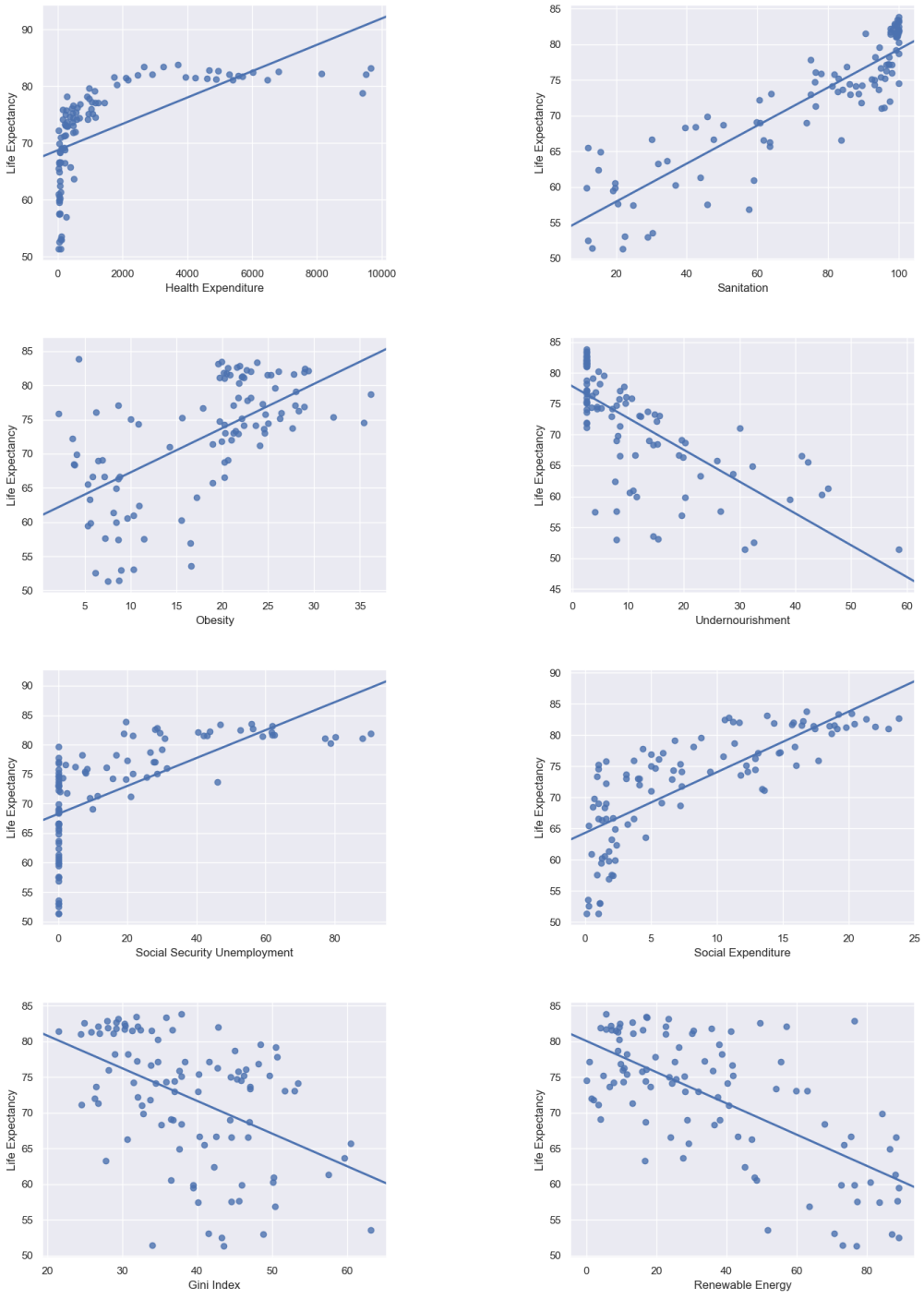

Figure 2.1 *Predictor variables (continuation)*

Based on the investigation of the data, we can see a non-linear relationship between Life Expectancy and some of the variables. Therefore, a preliminary transformation for the following variables is needed: GDP, Health Expenditure, Social Security Unemployment, and Social Expenditure.

As the trend in GDP, Health Expenditure, and Social Expenditure follows the log pattern, the logarithmic transformation was perform. Since the data of Social Security Unemployment contains "0" value, we have to find a different way from the log transformation to manage this predictor. One of the methods is to change this quantitative variable into a qualitative one ("0" when there is no social security unemployment, "1" otherwise).

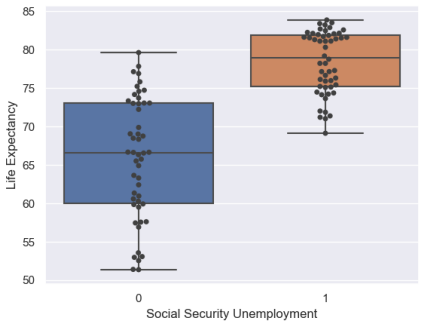

Figure 2.2 *Transformation of social security unemployment into a dummy variable*

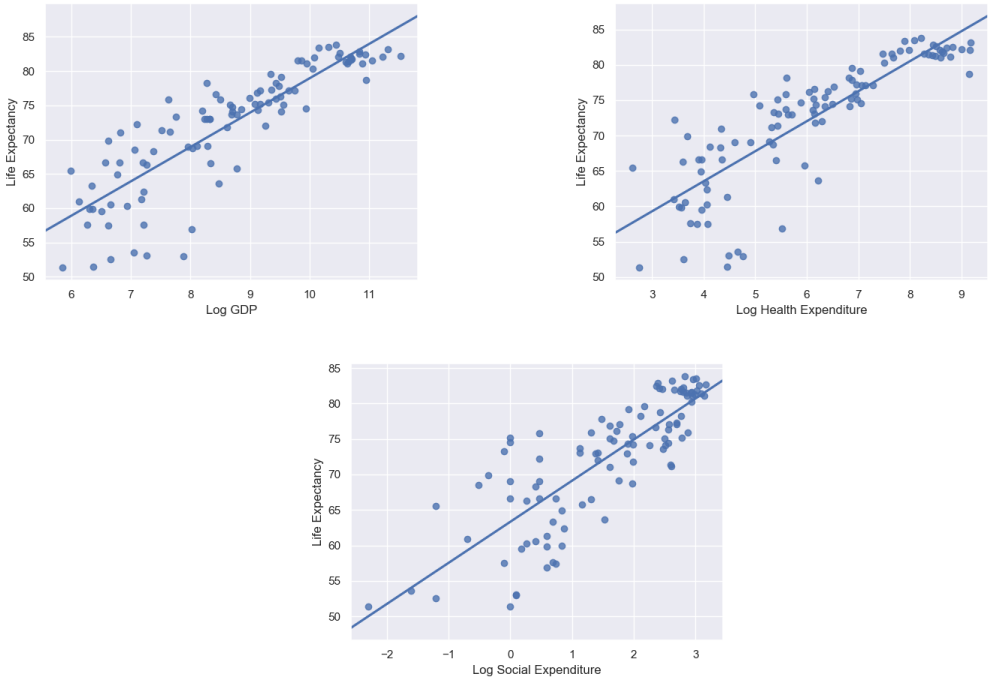

Figure 2.3 *Log-transformation*

3. Specification of the Model

Based on Figure 2.1, we can see that no relationship occurs between Life Expectancy and Forest. Therefore, we do not include this predictor in our model. The following regressors will constitute the preliminary model:

x_1 : Improved water source
x_2 : Renewable energy consumption
x_3 : Urban population
x_4 : log GDP per capita
x_5 : Services value added
x_6 : Exports of goods and services
x_7 : Developed economies
x_8 : Female labor force participation rate
x_9 : Log health expenditure per capita
x_{10} : Improved sanitation facilities
x_{11} : Obesity
x_{12} : Undernourishment
x_{13} : Social security unemployment
x_{14} : Log social protection expenditure

I have decided to add the interaction to the model : Obesity and Log social protection expenditure.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 \log x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 \log x_{i9} + \beta_{10} x_{i10} + \\ + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} \log x_{i14} + \beta_{15} x_{11} \times \log x_{14} + \epsilon_i$$

$i = 1,...,n$

Model assumptions:

1. The mean of response is a linear function of the x_i
2. The errors are independent
3. The errors are normally distributed
4. The errors have equal variance and mean zero

4. Estimation of the Appropriate Model

OLS Regression Results

Dep. Variable:	LifeExp	R-squared:	0.884
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	43.51
Date:	Thu, 10 Jan 2019	Prob (F-statistic):	1.20e-33
Time:	22:18:49	Log-Likelihood:	-256.31
No. Observations:	102	AIC:	544.6
Df Residuals:	86	BIC:	586.6
Df Model:	15		
Covariance Type:	nonrobust		

Table 4.1 *OLE regression results*

Test for significance of Regression:

$H_0 : \beta_1 = \ldots = \beta_{15} = 0$

$H_1 : \beta_j \neq 0$ for at least one j

$F_0 = 47.15$

p-value=0.000 < $\alpha = 0.05$

We reject the null hypothesis. We can conclude that there is a statistically significant linear association between the life expectancy at birth and at least one of our predictor variables.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	40.1129	6.627	6.053	0.000	26.938	53.287
WaterSource	0.0048	0.050	0.096	0.924	-0.095	0.105
RenewableEnergy	-0.0285	0.021	-1.355	0.179	-0.070	0.013
UrbanPop	0.0141	0.029	0.478	0.634	-0.044	0.072
Log_GDP	2.6702	1.290	2.070	0.041	0.106	5.235
Services	0.0251	0.046	0.546	0.587	-0.066	0.117
ExportsGoodsServices	-0.0133	0.013	-1.037	0.303	-0.039	0.012
DevelopedEcon	0.6419	1.332	0.482	0.631	-2.006	3.289
FemaleParticip	0.0609	0.031	1.948	0.055	-0.001	0.123
Log_Health_Expenditure	-1.0954	1.121	-0.977	0.331	-3.324	1.133
Sanitation	0.1823	0.030	6.105	0.000	0.123	0.242
Obesity	-0.2752	0.100	-2.762	0.007	-0.473	-0.077
Undernourishment	-0.0078	0.046	-0.168	0.867	-0.100	0.084
Dummy_Social_Unemployment	-2.2680	1.313	-1.728	0.088	-4.878	0.342
Log_Social_Expenditure	0.8399	0.916	0.917	0.362	-0.981	2.661
Log_Social_Expenditure:Obesity	0.0664	0.048	1.371	0.174	-0.030	0.163

Table 4.2 Parameter Estimates

The prediction equation is:

$\hat{y} = 40.112882 + 0.0048x_1 - 0.028468x_2 + 0.014056x_3 + 2.670223\log x_4 + 0.025143x_5 - 0.01332x_6$
$+0.641884x_7 + 0.060922x_8 - 1.095392\log x_9 + 0.182292x_{10} - 0.275188x_{11} - 0.007768x_{12}$
$-2.267975x_{13} + 0.839871\log x_{14} + 0.06638x_{11} \times \log x_{14}$

It appears from this analysis that only four following predictors are significant ($\alpha = 0.05$):

- Log GDP
- Female labor force participation rate
- Sanitation
- Obesity

This preliminary model explains 88% (R-Square=0.884) of variations in the life expectancy at birth. However, further tests of model adequacy are required. Moreover, from the investigation stage, we can presume that some of the estimated coefficients can be inflated by the existence of correlation among predictor variables (e.g. relationship between GDP and Health expenditure per capita).

5. Assessment of the Chosen Prediction Equation

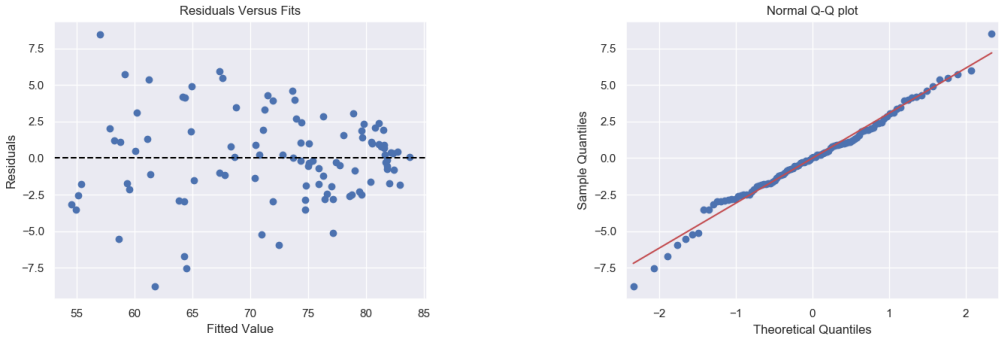

Figure 5.1 *Fit statistics*

37.634189	Lagrange multiplier statistic
0.001022	p-value
3.352235	f-value
0.000193	f p-value

Table 5.1 *B-P test*

The Breusch-Pagan test for constancy of error variance:

$H_0 : \gamma_1 = \ldots = \gamma_{15} = 0$ (constant variance)

$H_1 : \gamma_1 \neq \ldots \neq \gamma_{15} \neq 0$ for at least one j (non constant variance)

P-value=0.001 < $\alpha = 0.05$

We reject the null hypothesis. The variance of the residuals is not constant, which can be also proved by Figure 5.1.

Since the variance of the residuals is not constant, we should consider the transformation of Y values. The Box-Cox analysis suggests $\lambda = 3.88$, but we will continue with $\lambda = 4$.

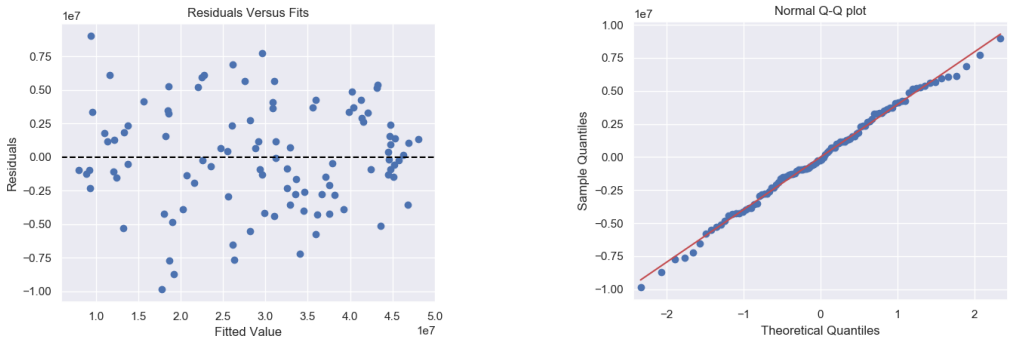

Figure 5.2 *Fit Statistics after the Box-Cox transformation*

Figure 5.2 presents the fit diagnostics after the Box-Cox transformation. We can see that variance of the residuals has been improved. Normality has been also improved.

24.203946	Lagrange multiplier statistic
0.061710	p-value
1.783757	f-value
0.050080	f p-value

Table 5.2 *B-P test after the Box-Cox transformation*

The Breusch-Pagan test for constancy of error variance: after the Box-Cox transformation:

$H_0 : \gamma_1 = \ldots = \gamma_{15} = 0$ (constant variance)

$H_1 : \gamma_1 \neq \ldots \neq \gamma_{15} \neq 0$ for at least one j (non constant variance)

P-value=0.06 > $\alpha = 0.05$

We fail to reject the null hypothesis. The variance of the residuals is constant.

We use the Shapiro-Wilk test for normality.

H_0 : the errors follow a normal distribution

H_1 : the errors do not follow a normal distribution

```
(0.9936488270759583, 0.9187743663787842)
```

P-value = 0.918 > $\alpha = 0.05$. We fail to reject the null hypothesis. The errors follow a normal distribution.

6. Outliers Diagnostics

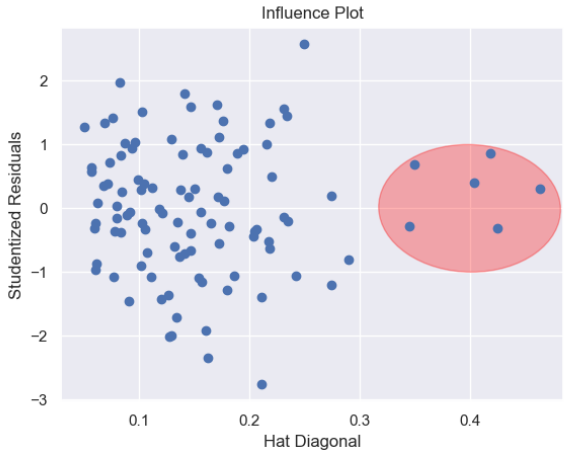

Figure 6.1 *Influence plot: studentized residuals vs. hat diagonal*

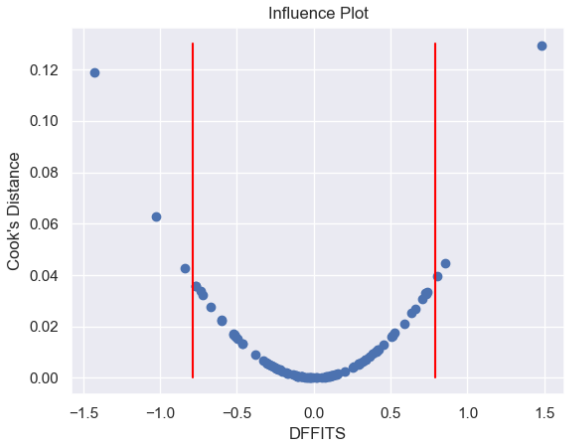

Figure 6.2 *Influence plot: Cook's distance vs. DFFITS*

Measure	Criteria	Observations
Outlier	|R|>3	none
h_{ii}	$\frac{2p}{n} = 0.314$	six observations
D_i	$\frac{4}{n} = 0.039$	none
|DFFITS|	$2\sqrt{\frac{p}{n}} = 0.792$	six observations

7. Multicollinearity Diagnostics

	VIF
Intercept	423.655194
WaterSource	4.334727
RenewableEnergy	3.080292
UrbanPop	3.892256
Log_GDP	36.495575
Services	2.857157
ExportsGoodsServices	1.401440
DevelopedEcon	3.619079
FemaleParticip	1.548174
Log_Health_Expenditure	36.932694
Sanitation	7.419658
Obesity	6.571195
Undernourishment	2.831238
Dummy_Social_Unemployment	4.154315
Log_Social_Expenditure	12.384251
Log_Social_Expenditure:Obesity	18.956394

Table 7.1 *Variance inflation factor*

Coefficients for Log GDP, LogHealth Expenditure and the interaction Log Social Expenditure & Obesity are highly inflated.

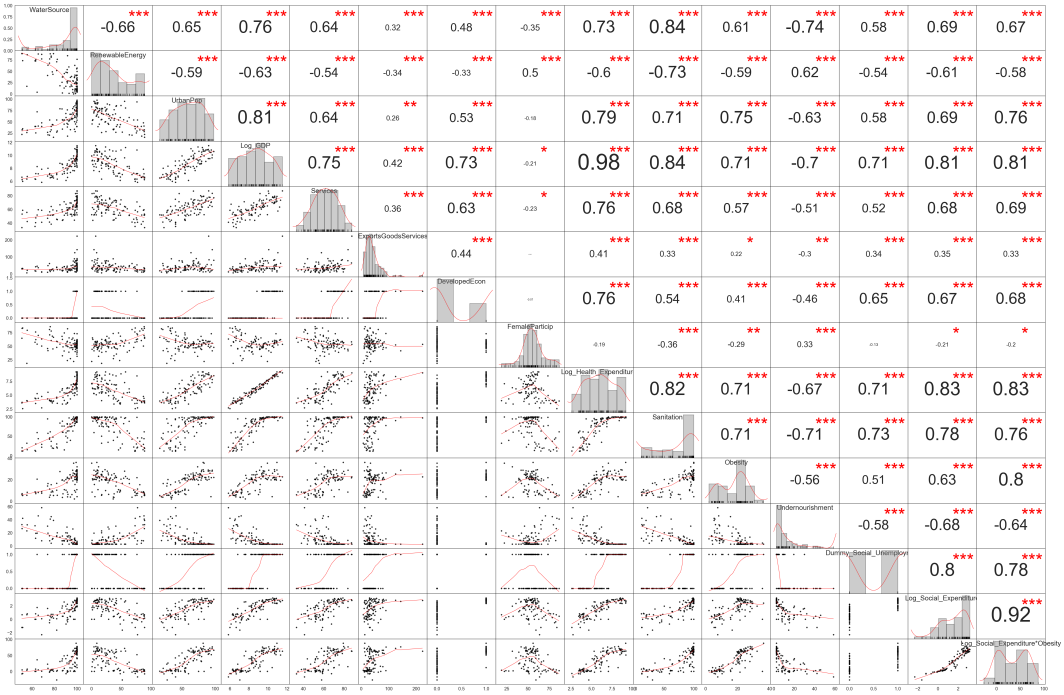

Figure 7.1 *Correlation matrix*

Based on the multicollinearity diagnostics, the following predictor have been removed:

- Log Social Expenditure & Obesity interaction
- Log GDP
- Log Social Expenditure

After the changes, the model includes the following variables:

- Improved water source, Renewable Energy, Urban Population, Services, Export Goods and Services, Developed Economy, Female Labor force participation rate, Log Health Expenditure, Sanitation, Obesity, Undernourishment, Dummy Social Unemployment

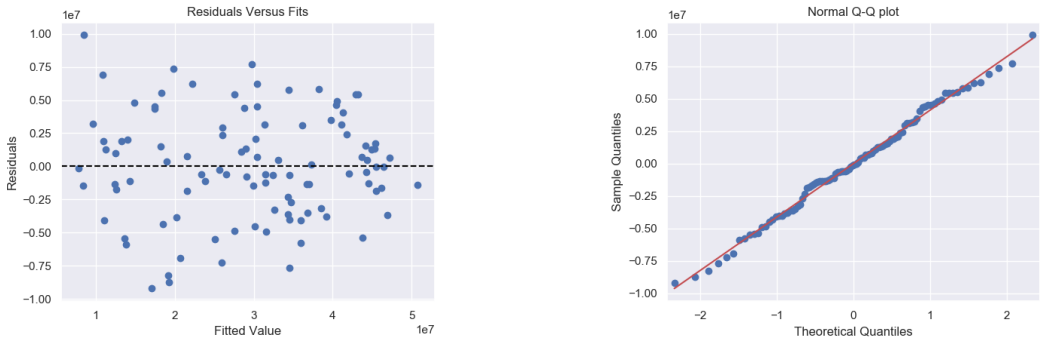

Figure 7.2 *Fit Diagnostics for a model after multicollinearity diagnostics*

We use the Shapiro-Wilk test for normality.

H_0 : the errors follow a normal distribution

H_1 : the errors do not follow a normal distribution

```
(0.9919253587722778, 0.8061215877532959)
```

P-value = 0.806 > $\alpha = 0.05$. We fail to reject the null hypothesis. The errors follow a normal distribution.

Based on Figure 7.2 and the Shapiro-Wilk test, we can conclude that the new model meets assumptions about constant variance and normality.

8. Selection of Variable Subset

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.173e+06	6.07e+06	-1.181	0.241	-1.92e+07	4.89e+06
WaterSource	6193.5792	6.51e+04	0.095	0.924	-1.23e+05	1.36e+05
RenewableEnergy	-1.687e+04	2.68e+04	-0.629	0.531	-7.01e+04	3.64e+04
UrbanPop	7.022e+04	3.72e+04	1.889	0.062	-3637.152	1.44e+05
Services	5.672e+04	5.89e+04	0.963	0.338	-6.04e+04	1.74e+05
ExportsGoodsServices	-1.824e+04	1.67e+04	-1.093	0.277	-5.14e+04	1.49e+04
DevelopedEcon	2.819e+06	1.67e+06	1.688	0.095	-4.98e+05	6.14e+06
FemaleParticip	5.402e+04	4.08e+04	1.326	0.188	-2.7e+04	1.35e+05
Log_Health_Expenditure	2.767e+06	6.69e+05	4.135	0.000	1.44e+06	4.1e+06
Sanitation	1.993e+05	3.82e+04	5.221	0.000	1.23e+05	2.75e+05
Obesity	-2.999e+05	8.56e+04	-3.504	0.001	-4.7e+05	-1.3e+05
Undernourishment	-2.709e+04	5.93e+04	-0.457	0.649	-1.45e+05	9.08e+04
Dummy_Social_Unemployment	4.008e+05	1.46e+06	0.275	0.784	-2.49e+06	3.29e+06

Table 8.1 *Parameter estimates after multicollinearity diagnostics*

With $\alpha = 0.05$, there are three following significant predictors:

- Log health expenditure
- Improved sanitation facilities
- Obesity

With $\alpha = 0.1$, there are five following significant predictors:

- Log health expenditure
- Improved sanitation facilities
- Obesity
- Urban population
- Developed Economies

Add Log_Health_Expenditure	with p-value 3.87997e-37
Add Sanitation	with p-value 1.8798e-09
Add Obesity	with p-value 0.0010338

Table 8.2 *Stepwise regression*

After the statistical analysis the final model includes:

- Log health expenditure
- Improved sanitation facilities
- Obesity
- Urban population
- Developed Economies

OLS Regression Results

Dep. Variable:	Transf_Y	R-squared:	0.889
Model:	OLS	Adj. R-squared:	0.883
Method:	Least Squares	F-statistic:	153.9
Date:	Sat, 19 Jan 2019	Prob (F-statistic):	3.20e-44
Time:	11:01:08	Log-Likelihood:	-1696.5
No. Observations:	102	AIC:	3405.
Df Residuals:	96	BIC:	3421.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.42e+06	2.05e+06	-2.161	0.033	-8.48e+06	-3.59e+05
UrbanPop	8.639e+04	3.51e+04	2.464	0.016	1.68e+04	1.56e+05
DevelopedEcon	2.837e+06	1.44e+06	1.970	0.052	-2.13e+04	5.69e+06
Log_Health_Expenditure	2.972e+06	6.33e+05	4.692	0.000	1.71e+06	4.23e+06
Sanitation	2.103e+05	2.61e+04	8.064	0.000	1.59e+05	2.62e+05
Obesity	-3.109e+05	8.28e+04	-3.755	0.000	-4.75e+05	-1.47e+05

Omnibus:	0.076	Durbin-Watson:	1.934
Prob(Omnibus):	0.963	Jarque-Bera (JB):	0.244
Skew:	0.005	Prob(JB):	0.885
Kurtosis:	2.761	Cond. No.	588.

Table 8.3 *Parameter estimates: model after variable selection*

With $\alpha = 0.1$, all regressors are significant.

$R^2 = 0.889$, which means that 89% of variation in the life expectancy at birth is explained by the model.

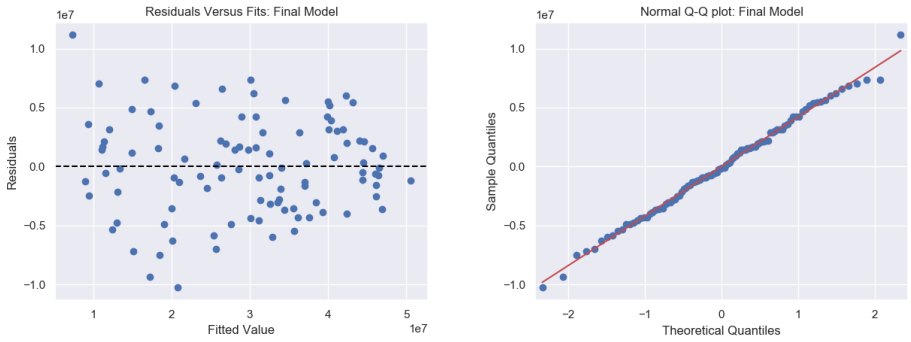

Figure 8.1 *Fit diagnostics for a model after variable selection*

10.696210	Lagrange multiplier statistic
0.057747	p-value
2.249274	f-value
0.055523	f p-value

Table 8.4 *B-P test: model after variable selection*

The Breusch-Pagan test for constancy of error variance: after the Box-Cox transformation:

$H_0 : \gamma_1 = \ldots = \gamma_6 = 0$ (constant variance)

$H_1 : \gamma_1 \neq \ldots \neq \gamma_6 \neq 0$ for at least one j (non constant variance)

P-value=0.057 > $\alpha = 0.05$

We fail to reject the null hypothesis. The variance of the residuals is constant.

We use the Shapiro-Wilk test for normality.

H_0 : the errors follow a normal distribution

H_1 : the errors do not follow a normal distribution

```
(0.9957419633865356, 0.9888380169868469)
```

P-value = 0.988 > $\alpha = 0.05$. We fail to reject the null hypothesis. The errors follow a normal distribution.

Based on Figure 8.1, B-P test and the Shapiro-Wilk test, we can conclude that the model meets assumptions about constant variance and normality.

9. Validation of the Regression Model

In order to validate the regression model, the dataset has been split into training (70%) and test (30%) sets.

```
Intercept:
 -4131028.1310854256
Coefficients:
 [  76916.97623905 3471361.93961773 3010507.0088227   180000.23103086
 -192791.29071341]
```

$$(\textbf{Life Expectancy})^4 = -4131028.131 + 76916.976 \times \textbf{Urban Population} + 3471361.93 \times \textbf{Developed Economy} + 3010507.008 \times \textbf{Log Health Expenditure} + 180000.231 \times \textbf{Sanitation} - 192791.29 \times \textbf{Obesity}$$

	Actual	Predicted
87	4.644019e+07	4.641366e+07
10	2.720625e+07	1.904125e+07
69	1.639923e+07	2.103710e+07
83	1.974689e+07	1.785213e+07
98	3.844270e+07	4.418138e+07
27	4.323888e+07	4.456772e+07
57	4.571995e+07	4.659344e+07
48	4.532525e+07	3.993610e+07
34	4.434702e+07	4.310932e+07
11	3.074884e+07	3.524177e+07
97	3.539404e+07	3.707151e+07
68	3.201384e+07	3.431819e+07
92	2.542188e+07	2.534451e+07
21	3.356826e+07	3.090367e+07
88	1.049123e+07	2.127383e+07
101	1.320304e+07	1.420390e+07
95	1.254075e+07	1.143586e+07
26	4.155933e+07	4.088205e+07
44	2.271793e+07	2.346249e+07
59	2.594732e+07	2.579436e+07
18	4.549741e+07	4.027069e+07
54	2.198130e+07	1.675216e+07
89	7.628190e+06	9.478568e+06
99	3.317153e+07	2.701547e+07
70	7.877344e+06	1.739784e+07
79	4.416704e+07	4.118296e+07
45	2.180279e+07	1.776151e+07
53	1.972820e+07	1.499310e+07
37	1.380724e+07	1.938304e+07
78	3.740549e+07	3.755222e+07
81	1.969887e+07	1.997797e+07

Table 9.1 *Fitted and predicted life expectation: test set*

The sum of squares of the prediction errors is $\sum e_i^2 =$ `578214015436609.1`

The corrected sum of squares of the responses in the prediction data set is:

$$SS_T = \sum y_i^2 - \frac{(\sum y_i)^2}{n} =$$ `2.2287751133930268e+16`

and the approximate R^2 for the prediction is:

$$R^2_{\text{pred}} = 1 - \frac{\sum e_i^2}{SS_T} =$$ `0.9740568704324614`

We may expect this model to explain about 97% of the variability in new data.

10. Conclusions

Research Question: What is the expected life expectancy at birth for a country where:

- Urban Population (in%): 34.277
- Developed Economy: 0
- Health Expenditure per capita (in USD): 30.833 (Log transformation needed: Log(30.833))
- Improved sanitation facilities (% of population with access):60.6
- Obesity (in %): 3.6

$$(\text{Life Expectancy})^4 = -4131028.131 + 76916.976 \times \text{Urban Population} + 3471361.93 \times \text{Developed Economy} + 3010507.008 \times \text{Log Health Expenditure} + 180000.231 \times \text{Sanitation} - 192791.29 \times \text{Obesity}$$

$$(\text{Life Expectancy})^4 = 19041250$$

$$(\text{Life Expectancy}) = (19041250)^{\frac{1}{4}} = 66.057$$